THESIS - KS142501

A COMPARISON OF MACHINE LEARNING TECHNIQUES: E-MAIL SPAM FILTERING FROM COMBINED SWAHILI AND ENGLISH EMAIL MESSAGES

Rashid Abdulla Omar
5216201701

SUPERVISOR
Dr. Ir. Aris Tjahyanto, M.Kom,

POSTGRADUATE PROGRAM
DEPARTMENT OF INFORMATION SYSTEM
FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2018

(This page is intentionally left blank)

APPROVAL SHEET

APPROVAL SHEET

A thesis submitted in partial fulfilment of the requirement for the degree of
Masters of Computer (M. Comp)
At
Institute of Technology Sepuluh Nopember, Surabaya, Indonesia 2018

By:
Rashid Abdulla Omar
NRP. 5216201701

Examination Date : January 4th, 2018
Graduation Period : March 2018

Approved by:

1. Dr. Ir. Aris Tjahyanto, M.Kom.
   NIP. 19650310 199102 1 001
   (Supervisor) 1

2. Dr.Eng. Febriliyan Samopa S.Kom, M.Kom.
   NIP. 19730219 199802 1 001
   (Examiner 1)

3. Nur Aini R., S.Kom M.Sc.Eng Ph.D.
   NIP. 19820120 200501 2 001
   (Examiner 2)

Dean
Faculty of Information and Communication Technology

Dr. Agus Zainal Arifin, S.Kom., M.Kom.
NIP. 19720809 199512 1 001

(This page is intentionally left blank)

A COMPARISON OF MACHINE LEARNING TECHNIQUES: E-MAIL SPAM FILTERING FROM COMBINED SWAHILI AND ENGLISH EMAIL MESSAGES

Student Name : RASHID ABDULLA OMAR
Student Number : 5216201701
Supervisor : Dr. Ir. Aris Tjahyanto, M.Kom.

ABSTRACT

The speed of technology change is faster now compared to the past ten to fifteen years. It changes the way people live and force them to use the latest devices to match with the speed. In communication perspectives nowadays, use of electronic mail (e-mail) for people who want to communicate with friends, companies or even the universities cannot be avoided. This makes it to be the most targeted by the spammer and hackers and other bad people who want to get the benefit by sending spam emails. The report shows that the amount of emails sent through the internet in a day can be more than 10 billion among these 45% are spams. The amount is not constant as sometimes it goes higher than what is noted here. This indicates clearly the magnitude of the problem and calls for the need for more efforts to be applied to reduce this amount and also minimize the effects from the spam messages.

Various measures have been taken to eliminate this problem. Once people used social methods, that is legislative means of control and now they are using technological methods which are more effective and timely in catching spams as these work by analyzing the messages content. In this paper we compare the performance of machine learning algorithms by doing the experiment for testing English language dataset, Swahili language dataset individual and combined two dataset to form one, and results from combined dataset compared them with the Gmail classifier. The classifiers which the researcher used are Naïve Bayes (NB), Sequential Minimal Optimization (SMO) and k-Nearest Neighbour (k-NN).

The results for combined dataset shows that SMO classifier lead the others by achieve 98.60% of accuracy, followed by k-NN classifier which has 97.20% accuracy, and Naïve Bayes classifier has 92.89% accuracy. From this result the researcher concludes that SMO classifier can work better in dataset that combined English and Swahili languages. In English dataset shows that SMO classifier leads other algorism, it achieved 97.51% of accuracy, followed by k-NN with average accuracy of 93.52% and the last but also good accuracy is Naïve Bayes that come with 87.78%. Swahili dataset Naïve Bayes lead others by getting 99.12% accuracy followed by SMO which has 98.69% and the last was k-NN which has 98.47%.

Key Words: Swahili, Gmail, Classifier, email, Naïve Bayes, SMO, k-NN

(This page is intentionally left blank)

PERBANDINGAN TEKNIK MACHINE LEARNING: PENYARINGAN E-MAIL SPAM DARI KOMBINASI PESAN E-MAIL BAHASA SWAHILI DAN BAHASA INGGRIS

Nama Mahasiswa	: RASHID ABDULLA OMAR
NRP	: 5216201701
Pembimbing	: Dr. Ir. Aris Tjahyanto, M.Kom.

ABSTRAK

Perubahan teknologi sekarang lebih cepat dibandingkan dengan sepuluh sampai lima belas tahun terakhir. Hal Ini mengubah cara hidup orang sehingga memaksa mereka untuk menggunakan perangkat terbaru yang sesuai dengan kecepatannya. Dalam perspektif komunikasi saat ini, penggunaan surat elektronik (e-mail) bagi orang yang ingin berkomunikasi dengan teman, perusahaan atau bahkan universitas tidak dapat dihindari. Hal tersebut menjadi target yang paling utama oleh spammer dan hacker dan orang jahat lainnya yang ingin mendapatkan keuntungan dengan mengirimkan email spam. Hasil Laporan menunjukkan bahwa jumlah email yang dikirim melalui internet dalam sehari bisa lebih dari 10 miliar dan 45% diantaranya adalah spam. Jumlahnya tidak konstan dan kadang naiknya lebih tinggi dari yang tercatat. Hal ini menunjukkan adanya masalah sehingga diperlukan upaya yang lebih besar untuk diterapkan dalam mengurangi jumlah spam dan meminimalkan dampak dari pesan spam.

Berbagai tindakan telah diambil untuk mengatasi masalah ini. ketika orang menggunakan metode sosial, itu merupakan alat kontrol legislatif dan sekarang mereka menggunakan metode teknologi yang lebih efektif dan tepat waktu dalam menangkap spam dengan menganalisis konten pesan. Dalam tulisan ini kami membandingkan kinerja pembelajaran mesin algoritma dengan melakukan percobaan untuk menguji dataset bahasa Inggris, kumpulan data bahasa Swahili dan menggabungkan dua dataset menjadi satu, hasil dari kumpulan data gabungan akan dibandingkan dengan pengelompokan Gmail. Pengelompokan yang digunakan peneliti adalah Naïve Bayes (NB), Sequential Minimal Optimization (SMO) dan k-Nearest Neighbor (k-NN).

Hasil gabungan dataset menunjukkan bahwa classifier SMO memiliki hasil yang lebih baik dibandingkan yang lain dengan hasil akurasi mencapai 98,60%, diikuti oleh classifier k-NN yang memiliki akurasi 97,20%, dan klasifikasi Naïve Bayes memiliki akurasi 92,89%. Dari hasil ini peneliti menyimpulkan bahwa classifier SMO dapat bekerja lebih baik dalam dataset yang menggabungkan bahasa Inggris dan bahasa Swahili. Dalam dataset bahasa Inggris menunjukkan bahwa pengelompokan SMO lebih baik dibandingkan dengan algorisme lainnya, dengan akurasi mencapai 97,51%, diikuti oleh k-NN dengan akurasi rata-rata 93,52% dan yang terakhir adalah Naïve Bayes dengan akurasi

87,78%. Dataset Swahili Naïve Bayes lebih baik daripada yang lain dengan akurasi 99,12% diikuti oleh Elective yang memiliki akurasi 98,69% dan yang terakhir adalah k-NN yang memiliki akurasi 98,47%.

Kata kunci: Swahili, Gmail, Classifier, email, Naïve Bayes, SMO, k-NN

DEDICATION

I would like to dedicate my thesis to my beloved mother Mrs. Maimuna Omar Ali, my wife Ashura and my daughter Ibtisam for supporting me all the time that I have been out of my country.

(This page is intentionally left blank)

ACKNOWLEDGEMENTS

I would like to say ALHAMDULILLAH for giving me ability to work on this thesis.

I would like to express my sincere thanks and gratitude to my supervisor Dr. Ir. Aris Tjahyanto whose invaluable assistance and guidance made the completion of this thesis possible.

My thanks also go to all my Institut Teknologi Sepuluh Nopember lecturers for their encouragement and moral support during my studies.

My greatest sincere thanks goes to my mother for her encouragement and moral support, and to my wife who courageously with support from her family endured the burden of looking after our daughter during my absence.

Thanks are also due to my friends at the Institute for making my stay in Indonesia enjoyable.

Finally, I wish to express my sincere gratitude to the Indonesian government especially KNB scholarship for their financial support that gave me the opportunity to study my masters in Indonesia, and the Ministry of Labour, Empowerment, Elders, Youth, Women and Children in Zanzibar for allowing me to pursue the course. Thanks all for enabling me to pursue my dreams.

(This page is intentionally left blank)

TABLE OF CONTENTS

(This page is intentionally left blank)

LIST OF PICTURES

(This page is intentionally left blank)

LIST OF TABLE

(This page is intentionally left blank)

CHAPTER 1
INTRODUCTION

In this chapter will be explained about the description of research starting from background, problem formulation, objectives, contribution of research, benefits of research and also limitation.

1.1 Background

The speed of technology change is faster now compared to the past ten to fifteen years. It changes the way people live and force them to use the latest devices to match with the speed of technological advancement. In communication perspectives nowadays, use of electronic mail (e-mail) for people who want to communicate with friends, companies or even the universities cannot be avoided. The traditional mail has many weaknesses including number of days it takes to be delivered, means of checking to know if it has been delivered or not, the time it takes for the sender to wait for a reply, and is also unreliable. So the importance of using e-mail as the main means of communication cannot be underestimated. This makes emails to be the most targeted by bad guys especially hackers to get the benefit from the email users.

Many agencies tried to investigate this issue in-order to know how many email accounts were affected and how the number increases so that they could predict growth in the future. THE RADICATI GROUP, INC in summary of the Email Statistics Report, 2015-2019 reported that, they are expecting the total number of email accounts worldwide to increase from nearly 2.6 billion in 2015 to over 2.9 billion by the end of 2019. This represents an average annual growth rate of about 3% over the next four years". The amount of email messages will also increase at the end of 2017, (N pérez-díaz et. al. 2016). The total usage of sending and receiving email statistics shows that it increases each year, and for the year 2015 there was an average increase of 538.1 million messages per day. The statistics shows that there has been an increase of 5% since 2010. In general, emails trends is expected to increase more in the coming years, where the business

emails are specifically expected to increase to over 132 billion more in sending and receiving email messages per day by the end of 2017.

As mentioned earlier in the first paragraph, the use of electronic mail has increased greatly as people can now send and receive emails where ever they are, be it at home or even travelling around the globe. This has been made possible the use of smart phones. Those who use smart phones which have android operating system for example, are forced to have at least one Gmail account, that account help them to access other facilities including download applications in their phones. So, by being forced to have an account, the owners will use the emails some of them without proper prior knowledge on good usage and what danger there is when using such appliances inappropriately. The possibility to be among the victims is more than 70% because they will use their accounts carelessly by opening the emails which contain viruses that can crash their phones or submit their personal information to bad guys and misuse them. Nowadays there are social engineers and spammers that pretend to be someone you know, who ask users to follow the link which will lead them to their website (phishing sites).

This problem can be reduced, if not solved, by the classifier to separate the emails into different folders. In the classification of emails the message can be classified into two groups, the one which is legitimate, also called as non-spam, which means the message is harmless, and the second type is bulk e-mail or unsolicited e-mail message, also known as spam. Unsolicited messages are normally distributed by using bulk-mailers and address lists harvested from different web pages or in news group archives. The message content varies significantly, some are vacation advertisements to get-rich schemes, some of them are the advertisement of products like Viagra and others can also come from the service companies.

The common feature of these messages is that they usually have little interest to the majority of the recipients. In some cases, they may even be harmful to the recipient, and some spam messages advertising, pornographic sites. It will not check the recipient's age, and possibly be sent and read by children. Un-harm spams sometimes just waste your time and bandwidth, especially for those who

use dial-up connections. Apart from this, spam e-mail also costs money to users. The reports from spamlaws.com said that spammer from all over the world use their accounts for sending 14.5 billion messages daily which is almost 45% of all emails sent. Some research companies estimate that spam email makes up an even greater portion of global emails. Some reports put the figure to be up to 73%. The country which is number one on the ranking of the spam or unwanted email senders and recipients is the United States, followed closely by South Korea. These countries are the largest spam messages distributors in the world. The report also shows that advertising-related email type of spam is leading when you compare with other types of emails. This type of spam accounts for approximately 36% of all spam messages. The second most common category of spam is on adult-related subjects and makes up roughly 31.7% of all spam. Unwanted emails related to financial matters are the third most popular form of spam, at 26.5%. Surprisingly, scams and fraud comprise only 2.5% of all spam email; however, identity theft which is known as phishing makes up 73% of this figure (2.5%), the remaining shares with others like botnet etc.

According to the Anti-Phishing Working Group (APWG) reported in the fourth quarter of the year 2015 and also 4th quarter of the year 2016, it shows that the email phishing is not fixed because in some months it goes higher but in some months it becomes low which means it is not predictable although in general email phishing is still a big threat and a challenging one. The number of unique phishing e-mails reported in 2015 campaigns received by APWG from consumers in the fourth quarter shows that in October it was 48,114, November 44,575, and December 65,885. The sum of emails which have phishing attacks observed in the fourth quarter was 158,574. This shows the increase of over 21,000 phishing sites detected during the holiday season. In 2016, during the 4th quarter, the record for October was 51,153, November 64,324 and December 95,555.

Protection is needed in order to reduce the damage that is caused by spam emails. Spammers are working hard to organize criminal activities, illegal trafficking of goods and services in the stock market fraud, wire fraud, identity theft and hijacking using computers. This is very costly in business when you want to respond on request of your customer (Thiago S. Guzella, 2009). The cost

caused by spam in terms of lost productivity in the USA has reached USD21.58 billion per year and worldwide USD50 billion (Tiago A. 2011). The individuals incur 10% cost in spam email according to (Ion Androutsopoulos 2000) this cost including the waste of bandwidth for the dialup connections.

Due to the seriousness of the issue in hand, a lot of researches have been done using dataset in English, Arabic and Chines languages. It is unfortunate though that there is no research done using Swahili language. Swahili (Kiswahili) is a language that is widely spoken in all East African countries of Tanzania, Kenya and Uganda. Countries such as Rwanda and Burundi who have recently joined the East African Community as well as other neighboring countries including Democratic Republic of Congo, Malawi and Mozambique have also started using the language to ease communication especially in the area of trade. The Swahili is very complex in terms of structure and the addition of suffix, this make the verb to be a complete sentence (“anakimbilia mpira” he is running for the ball). The applied suffix on Swahili verbs has long posed an analytical problem, the basic meaning of this suffix has to do with "directing the action against something" (Port, R. F. 1981). Also the negation in Swahili is different when you compare with English language, this part is very complicated because it does not have the specific words and position of the word for refusal/opposite (Contini-Morava, E. 2012). In this way, it is difficult to know whether the messages sent through the email using Swahili are spam or not, contrary to the ones sent through the English language whose vocabulary is largely standard.

The Swahili people use technology and have so many researches, but until now there is no research on email spam which has been conducted using Swahili language. It makes near impossible to get the dataset that is written in Swahili so that force us to create our own dataset in Swahili language, to be used in this research. There are some challenges though, the first one is time to collect all emails that are written in Swahili language.

The ways classifiers algorithm is used help to reduce the impact of the spam. In this thesis, we will check the ability of google mail (Gmail) classifier to see how accurate it is in detecting spam emails because it has been found out that although it works well but it has some weaknesses. This has led to the decision to

tackle this problem in this thesis. To begin with, we went through many email addresses which are @gmail.com and two more were also created specifically for use in this thesis. It was found out that some spam emails were actually in the inbox. This should not be the case as they were supposed to be at junk/spam. It was also observed that some non-spam emails were in spam folder, and sometimes names of the email addresses confused the Gmail classifier, this led to the emails to be put in wrong or inappropriate folders. At a later stage, calculations were done manually from one of our Gmail accounts, to determine the performance of the emails received by using confusion matrix. The result showed that 86.26% of the emails were correctly classified, while those wrongly classified were 13.74%. There is a possibility for this to be improved by two to four percent.

The solution of this problem includes selection of classifiers that achieve expectations that the researcher have. In recent years, any researches have been done in recent years on text categorization which suggested many methods, among them are Naïve Bayes which was singled out as an effective method to construct automatically anti-spam filtering with good performance (Ion Androutsopoulos 2000), (Sahami 1998), (Daelemans et. Al. 1999), (Koutsias, J., et. al. 2000, July). All these papers compared Naïve Bayesian algorithm and other algorithms NB come with best result. Also in other researches of (Yu, B., & Xu, Z. B. 2008) shows the Naïve Bayes and Support Vector Machine both perform well. The researchers (Hmeidi, I., et. al. 2015) using Arabic dataset tried to compare the classifiers which are Naïve Bayes, Support Vector Machine, Decision Tree, Decision Table, and K-Nearest Neighbour (KNN). The results showed that Support Vector Machine leave behind all the other classifiers. Among the ones that were suggested by the researchers, the author choose three - Naïve Bayes, Support Vector Machine and K-Nearest Neighbour.

1.2 Problem Formulation

The background of the thesis states that spam emails are increasing, according to the researchers which were cited above. So, the problem formulation is as follow:

1. How can the features be extracted in such a way that the classifiers' work could be simplified in order to increase accuracy?
2. What would be the performance of classifier if the dataset is a combination of two languages (Swahili Language and English Language)?

1.3 Research Objectives

In this research objectives are divided in to two parts, one which is the main/general and the ones which are specific in the classifying.

1.3.1 General Objective

The main intention of this thesis is to compare by increasing the accuracy of the classifier algorithm that Gmail used to classify the personal emails by replacing with other machine learning classifier which can work better than the Gmail with a high accuracy.

1.3.2 Specific Objective

To reduce dimensionality in such a way that will lead to get better performance for classifiers. To identify the best classifier in email classifiers which can be compared with the Gmail classifier. To find the classifier that can process the dataset within appropriate time.

1.4 Contribution

Due to the seriousness of the issue at hand, a lot of researches have been done using dataset in English language (Zhang, I, Zhu, J, & Yao, T. 2014). The researchers tested the three English dataset with one Chinese dataset. On the other hand other researchers used the Arabic language for the same purpose. These are (El-Halees, A. 2009), (Hayati, P., & Potdar, V. 2008) (Al-Harbi, S. at al 2008), (Khorsheed, M. S, & Al-Thubaity, A. O. 2013) and (Hmeidi, I. at al 2015) in these five researches they tested algorithms which are written in Arabic language. Chinese language has also been used, where (Dong, J., Cao, H., Liu, P., & Ren, I. , 2006, October). It is unfortunate therefore that there is no research

done using Swahili language and also no dataset that contain Swahili emails. The main contribution in this research is:

- To have dataset that combine two languages (Swahili and English)
- To test the Algorithms performance on English and Swahili dataset.

1.5 Benefits of research

The research benefits will be in two different dimensions, theoretical benefits and practical benefit of our research.

1.5.1 Theoretical benefits

The theoretical benefits of our research includes firstly enhancing the potential knowledge base of the text categorization, especially on email classification research. This means that other researchers will have the foundation to develop research beyond this research. Secondly, the researchers can find a gaps from our research finding and so conduct their own research to improve the result.

1.5.2 Practical benefits

Practical benefits of our research among many include the awareness creation to email users including individuals, researchers, Practitioners, policy makers and also the managers of the companies who are the decisions makers to provide the support needed especially in the IT Department in order to keep the computers and other related machines free from spam and other malicious programs. Also the engineers can use research findings through practical applications to improve products and system implementation.

1.6 Limitations of Research

The research will only test one to three emails and the dataset which the author created might be of lower standard compared to the dataset that have been created by the professionals who have been making dataset for many years.

(This page is intentionally left blank)

CHAPTER 2
LITERATURE REVIEW

In this chapter, we tried to explain briefly according to some researchers about document sorting and text categorization. Our focus will be on e-mail classification (Spam and Legitimate) to see what the effect was in the past few years, and how the hackers use Spam for their benefits. Also, observations will be made on how ant-hacking guys initiated efforts for fighting with spammers by using a classifier algorithms with the aim of minimizing the effect which posed by the hacker through e-mails.

2.1 Application of text categorization

The categorization of text has three applications which are commonly used. The first one is text indexing where a text document is used to assign keywords from a controlled vocabulary (Sebastiani, F. 2002). The second one is web page categorization, which is the process of assigning a web page label to one or more categories; while the third one is document sorting and text filtering which process the incoming document by sorting and filtering them according to the categories like sales and personal, spam or legitimate email. In this research our focus will be on text filtering based on email filtering.

2.2 Email group (Spam and Non-Spam)

In this section we will only focus more on spam rather than non-spam messages or legitimate messages. A legitimate message can be defined as a message that comes from the source which the recipient knows or expects to receive a message from them. This will not create any doubt because he/she knows the sander(s), and also the message itself will not contain any spam content. Sometimes this is not the case because spammers use the addresses which the recipient knows to send spam messages. The non-spam messages must not be harmful this is the main point, but in the body, subject or address the content is not specific in terms of words used.

The Unsolicited Bulk Emails (UBE) or spam are the messages which mainly come from unknown sources, although sometimes they may come from the known sander address but the content will be deferent. Unsolicited means the recipient is not expecting to receive any email. Bulk means a message is sent out as part of a large number of messages with all having substantively identical content at a reduced rate (Spamhaus Project 2017). Spam can come in the form of an advertisement which do not have any harm as its objective is only to advertise and promote a product. This is considered as time wasting. On the other hand there are spam messages which intend to cause the damage to recipient or to his/her network infrastructure. There are so many groups of spam, among them are phishing and social engineering. Although these contribute to a small amount, yet they are so dangerous for the email users.

2.3 Swahili Language

The Swahili language (Kiswahili) is widely spoken in all East African countries where 99% of people living in Tanzania, 87% in Kenya and 85% in Uganda use the language. Countries such as Rwanda 28% and Burundi 55% who have recently joined the East African Community has 28% and 55% Swahili speakers respectively (Gbogboti 2012). The neighboring country, the Democratic Republic of Congo has 48% of its population who speak the language. Other neighboring countries including Malawi and Mozambique have also started using the language to ease communication especially in the area of trade. Swahili, that originates from Arabic and Bantu (African) language has the same alphabet as English language but uses a different combination in word formation. Over the years, new words from different ethnic languages have been added and now are used in day to day communication including through correspondence using the electronic media.

In some cases it makes it difficult for someone who is not conversant with such words to fully understand the messages. In this way, it is difficult to know whether the messages sent through the email using Swahili are spam or not, contrary to the ones sent through the English language whose vocabulary is largely standard. Swahili language has three different types of sentences: a simple

sentence that consists of a single clause; a complex sentence that consists of one main clause and at least one subordinate clause which obligatorily follows the main clause, and a compound sentence that consists of at least two main clauses joined by a coordinating conjunction. In terms of word order the Swahili language has fixed order Subject Verb Object at the sentence this means the subject come first before verb and object "Ally anacheza mpira" means "Ally is playing football" (a-na-cheza 'he-present-play').

The Swahili language in the verbs they use suffix, this make the verb to be a complete sentence ("*anakimbilia* mpira" *he is running for* the ball) when translate in English can come with a phrase with more than three words. The applied suffix on Swahili verbs has long posed an analytical problem, the basic meaning of this suffix has to do with "directing the action against something" (Port, R. F. 1981).

Example 1:

Pig-a	'strike'	pig-i-a
Omb-a	'pray'	omb-e-a
Chuku-a	'take'	chuku-li-a

Example 2a:

alikata nyama " he cut meat"

a-li-kat-a
S(he)-Past-cut-Indic

aliikata nyama " he cut the meat"

In the example 2b the subject and object pronouns are 'he/she' and 'me' correspondingly, and the verb is suffixed with IE. Meanwhile the implication of the sentence is 'he cut meat for me', actually IE adds the role of a beneficiary or indirect object that is played by the first person singular pronoun in the object prefix.

Example 2b:

alinikatia nyama "he cut the meat for me"

a-li-ni-kat-ia
S(he)-Past-O(me)cut-IE

Nilikatiwa nyama “I had meat cut for me (by him)”

Nitakupokelea zawadi “I will accept the gift for you”

Ni-ta-ku-poke-lea	Ni-li-m-shindili-lia
S(I)-future(will)-Object(you)-accept	S(I)-past-Object(him)-pack

Nilimshindililia majani “I packed down the leaves for him” the word ‘shindilia’ means to press down.

The issue of negation in Swahili language is different when you compare with English language, this part is very complicated it does not have the specific words and the position for refusal/opposite sometime can be at the beginning like in the example 4a or in the middle of the word in example 4b (Contini-Morava, E. 2012). In the example show the sentence that use negation.

Example 4:

a. Bwana Mussa, niliduwaa **nisiwe** na lakusema “Mr. Msa, I was dumbfounded, I didn’t have (neg. subjunctive) [anything] to say”

ni-	*si-*	*w-*	*e*
1sgSubj	Neg2	be	subjunctive

b. Mahmoud alisema **hatutawaona** “ Mahmoud said we will not see them”

Preinitial (Neg1)	Initial Subj.	Postinitial TAM	Prerad. (Obj.)	Radical VStem	Final TAM
(*ha-*)	*tu-*	*ta-*	*wa-*	*on-*	*-a*
	1pl.	Future	3pl.	see	indic.

In technological terms the Swahili language use to borrow the same words but pronounced in Swahili way like spam “spamu”, computer “kompyuta”, virus “kirusi” viruses “virusi”, but also they have their own which are sometimes not know among Swahili, for example peoples email “barua pepe”, password “nywila” (Petzell, M., 2005).

2.4 Spammer motivations

The things which motivate the spammer to continue spreading spam messages include income generation using google AdSense™. People nowadays

make a lot of money through advertisement using the websites, google AdSense™ an organization that pays a lot of money on that. Spammers exploit the services by generating copied (synthetic) content and then monetize (earn revenue from) it from the AdSense™. Some spammers rank their websites incorporate with search engine optimization techniques to get their website a higher rank, with outcomes in extra traffic and consequently more revenue via advertising. When more users access/visit a website, will give the credit to that site and also increase the rank for a web site. By Promoting Products and Services, the spammer get paid by the company which they work with in order to advertise their product. The reasons which are mentioned above means some spam are not intended to harm or intrude the user's privacy or security, it is just a waste of the bandwidth and time for the recipient. Meanwhile there are spammer are who motivated by stealing of someone's confidential information such as bank account, PIN, username and passwords and also target to destroy the network or make it busy (phishing and botnet) (Hayati, P., & Potdar, V. 2008).

2.5 The Damage Caused by Spam

The protection needed in order to reduce the damage that is caused by spam which include network bandwidth wasted and time spent by distinguishing the spam and non-spam is very costly in business when you want to respond on request of your customer (Thiago S. Guzella, 2009). The cost caused by spam in terms of lost productivity in the USA has reached USD21.58 billion per year and worldwide USD50 billion (Tiago A. 2011). The individual incur 10% cost in spam email according to (Ion Androutsopoulos 2000**).** This includes the use of gigabit which you pay for your ISP or telecommunication company. Spam can be a malware carrier this means that some spam carries email attachments that if opened can infect your computer with viruses or spyware. The spammer use some viruses that are engineered to install spam-sending software on a victim's computer. This can make you lose your internet services because if the internet services provider (ISP) see your computer as a source of spam, ISP will cut their service from that computer.

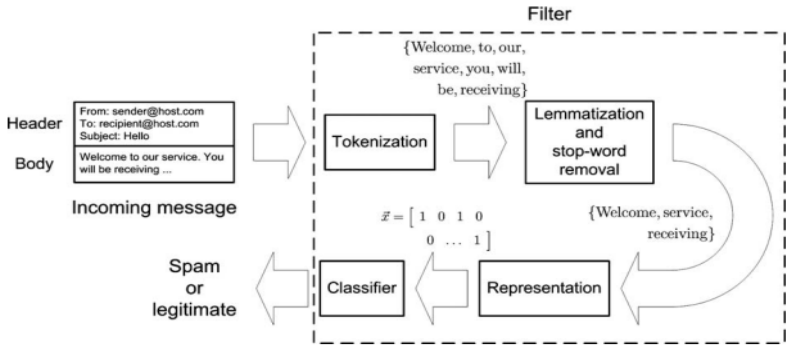
2.6 Fighting Spammer approaches

The fighting against spammer is very challenging and very hard as written in many spam classification researches. There are many ways to defeat spammer. Some countries use social ways in fighting against the spammer by posing a legislative means of control, although some researchers have reported that this method has least effect when it comes to the fight against spammer for many reasons citing the way in which the technology changes quickly while laws take time be repealed/amended and so not preventive, it's just an action taken after the event. The law succeed only in regulating service providers (ISP's) and other entities that manage internet but not spammers because they just use the resources without any permission and have many tricks to carry out their tasks (Hoanca, B. 2006).

Other approach used in fighting against spammers is technological which use the filtering techniques (spam filter). This technique is based on analyzing of the messages content (sender and body) and other information which can help to identify the messages if it is spam or not immediately before even they cause any harm. After identifying messages that contain spam, the action that follows depend on the setting which are applied by the filter itself. There are client-side filter which usually send the spam messages to a special folder (spam folder) to make the identification easier. Some filters also operates in a mail server and these will take a different action either by deleting the message or just label it as a spam. There are also some machines which use the collaborative settings. This means that the filters are running in special machines to identify and then share information to other machine on the messages received which will help to improve their performance. (Thiago S. et. la 2009).

The figure bellow shows the main steps taken in spam filtering. When the message is received the first step in the process is to extract the words in message body (tokenization). This is followed by the second step which is to transform the words to its base form (lemmatization (e.g., ‘‘extracting” to ‘‘extract”)) then, the stop-words removal take palace by removing the words which occur often in many messages (e.g., ‘‘to”, ‘‘a”, ‘‘for”); and finally the

representation change the messages in the format which machine learning algorithm can use for classification (Thiago S. et. la 2009).



Picture 2. 1: Email filtering

2.7 Classification Techniques

The classification techniques are many. While there are some techniques which were once used and now are no longer functional, there are others though which can still be applied. Many papers presented in various settings show that the most popular of email classification techniques that is being used in text classification include naïve Bayes, rule learners, and support vector machines. Most of these techniques examine or concentrate on the words that means text in the message headers and body to predict its folder classification (which folder the message belongs to).

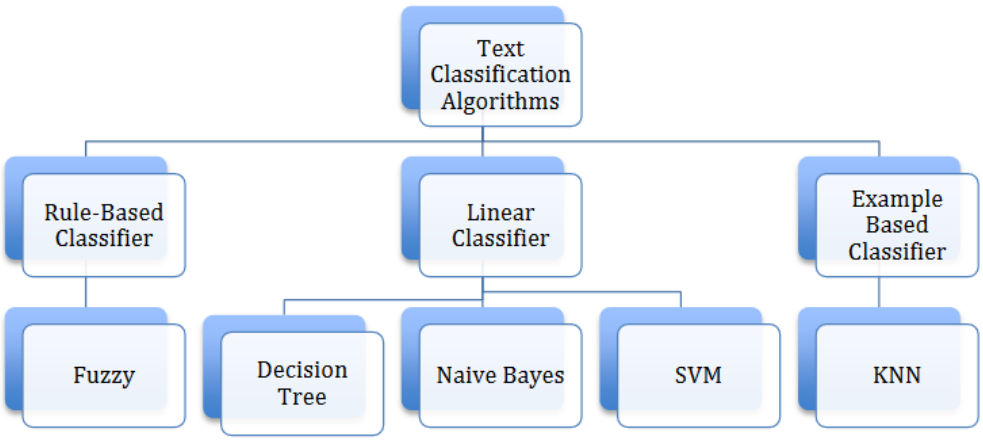
2.7.1 Whitelist and Blacklist

The earliest popular approach was based on blacklists and whitelists. A blacklist is a list of email addresses (senders) or IP addresses whose emails want to be blocked from being sent to the recipients (domain). This can be done by administrator or user themselves by using the blocked sender setting, this will tell the filter automatically to send those messages to trash. A whitelist method is working as the opposite of blacklist. A whitelist allows only those messages or addresses who are saved on the list to get through. Meanwhile spammers practically always spoof the “From:” field of spam messages, typically blacklists

concentrate on IP addresses without taking consideration on the email addresses. For other incoming messages from the senders which do not appeared on the lists, content-based filters might be practically applied so that the two approaches can complement one another (Lam, H. Y., & Yeung, D. Y. 2007).

2.8 Text Classification Algorithms

The text classification algorithms according to Phadke, S. G. (2015), is divided into three parts which are rule-based classifier. This type of algorithms works according to the set of instructions or rules that are set to classify data. For example based classifier rely on directly computing the similarities between the document to be classified and the training document, and linear classifier.



Picture 2. 2: Taxonomy Structure of Text Classification Algorithms

2.8.1 Naive Bayes Classifier

The machine learning uses voluminous classifier techniques one among them is Naïve Bayes Classifier technique. This is a simple but effective working tool used in several applications of information processing and gives good results. The classifier based on Bayesian theorem and it works well mainly when the dimensionality of the contributions are high. NB have been applied in many applications of information processing, example natural language processing, text categorization and information retrieval. NB is competent when inputs of

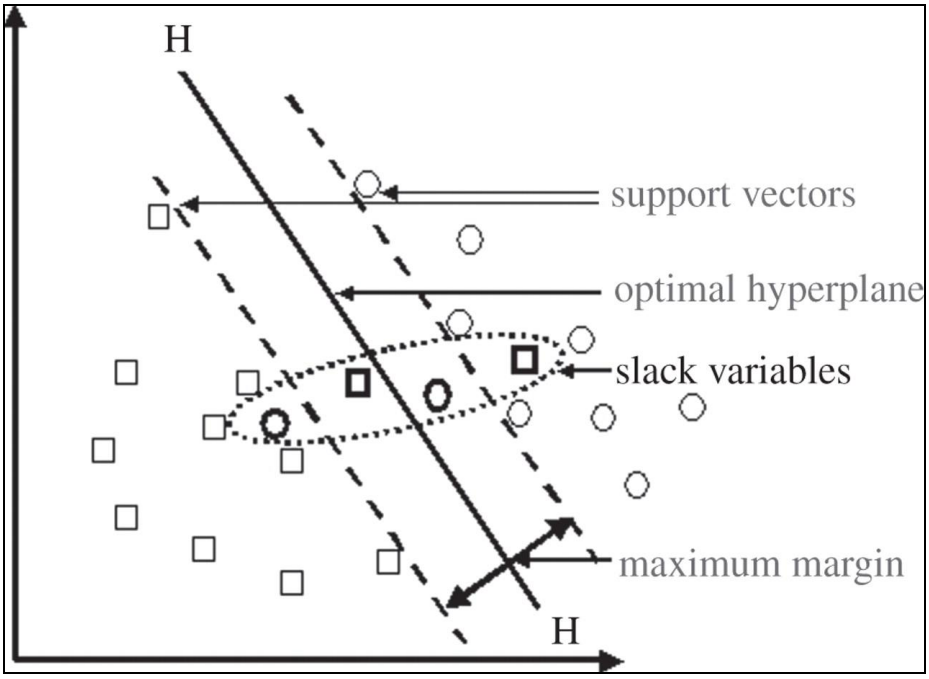
dimensionality is high. The assumption of Naïve Bayes classifiers is that effectiveness of a value in a certain class is independent of the values of other variable. Naïve-Bayes also computes conditional likelihoods of the classes given the instance and choices the class with the highest posterior. In supervised learning setting Naïve Bayes classifiers is more capable and more efficient to be trained. **Bayes** rules applied to documents and classes, this class can be ham or spam in email context (Tretyakov, K. 2004, May). Formula 1. Show probability P of document d in a given class c equal to P(d | c) P(c) dived by P(d).

$$P(c|d) = P(d|c)\ P(c)/P(d) \qquad (2.1)$$

2.8.2 Support Vector Machines

The support vector machine (SVM) is a supervised learning approach that is used for classification, and it generates mapping functions (input-output) from a set of labeled training dataset. The mapping function either can be (1) a classification function, that is a type of the input data, or (2) a regression function. For classification functions, nonlinear kernels are often used to convert input data to a high-dimensional feature space (Wang, L. (Ed.). 2005). That input data developed is more independent compared to the original input space. What results after that are creation of maximum-margin hyperplanes. The class produced therefore depends on only a sub-set of the training data near the class margins.

Picture 2.3 shows how SVM works. Hyperplane is a line that separate objects of different classes. Hyperplane can be one or more but the best is the one that leaves maximum margin from the nearest points of each class to the separating hyperplane (Youn, S., & McLeod, D. 2007). Support vectors are the coordinates of the training examples, which are closest to the classifying hyperplane. Slack variables are the variable that belong to one class but can be found on the other side of hyperplane.



Picture 2. 3: SVM classifier structure

2.8.3 Sequential Minimal Optimization

Sequential Minimal Optimization 'SMO' classifier is among algorithm that are intend to solve some weaknesses that researcher find in support vector machines (Zeng, Z. Q., et. al. 2008, November). SVM is too slow especially when trying to execute large scale dataset, so sequential minimal optimization and stochastic gradient descent has come to improve SVM weaknesses (Platt, J. 1998). SMO was discovered soon after SVM was operational, its write everything in terms of kernel, and it optimizes two variables at a time. SMO have the ability of execute very large dataset without requiring extra matrices storage, and it does not invoke any repetition of routine number for every sub-problem. SVM is not capable of doing so because it must have extra matrices storage, and so this is the main weaknesses of SVM. This problem called quadratic programming 'QP'.

2.8.4 K-Nearest Neighbors

The k-nearest neighbor (k-NN) classifier is considered by some as an example-based classifier. This means that when the document is used as a training document, will take it as it is and use it for comparison rather than an

explicit category representation. The membership of class in k-NN is allocated to a vector but not assigned a vector to a specific class. This has some benefits like there being no random assignments are made by the classifier (Keller, J. M., et. al. 1985).

The categorization can be done when the new documents have to be classified, where comparable document, that is neighbors, are discovered. If the document is assigned to a category then the new document will also be assigned to that category. The nearest neighbors by using the traditional indexing can be found easily and quickly. To decide which group does a message falls into, that is whether legitimate or spam, we consider the messages classes that reside near to it. We conclude by saying that the comparison of the vectors are real time process.

2.8.5 Gmail Filter

Gmail is among well-known webmail service provider, Gmail was introduced to the public in 2004. To agree if an email is a spam or not, numerous rules must be applied for every incoming email that reach in Google's data centers. These rules are able to detect general spams. Gmail is used to filter email messages automatically, the technique used by Gmail filter is unidentified, Mojdeh, M. (2012). The author thinks that it combine many processes including network authentication, blacklists, and others. Also it filters by giving user a choice whether the email messages received are spam or non-spam, by giving him an option. The user can decide if that email message is spam or not by selecting the email message and click send to spam, and is broadcasted as learning from this feedback, that means user's choice will be remembered by the system. Gmail declared that with the spam occurrence of seventy percent (70%) in the year 2007, and the user's testified rate of a smaller amount cannot reach 1% as spam in their inbox (Mojdeh, M. 2012).

2.9 Text categorization approaches

There are two main approaches in text categorization which are the Knowledge Engineering approach in which the expert's knowledge about the categories is directly encoded into the system declaratively or in the form of

procedural classification rules, and Machine Learning (Konstantin Tretyakov 2004).

2.9.1 Machine learning

Machine learning is the way of simplifying tasks by studying the computer algorithms and learning how to do the task in an easy way. We will use machine learning in this thesis and apply some classifier algorithms. Sometimes we can use or MATLAB by writing the code from the scratch which take more time compared with using WEKA or other machine learning tools. The learning acquired is always constructed on some sort of observations and/or data, plan to do better in the future based on the experience we got earlier.

2.9.2 Waikato Environment for Knowledge Analysis (Weka)

Weka is the data mining tools which can be used for classifying and clustering information. It's a pool of algorithm from machine-learning, among them are classification, regression, clustering, and association rules to complete the data mining tasks. The interface can link with email information to gather the information for pre-processing then generate the coaching and take a look at data sets and then to convert each set into rail format. We have a tendency to pass coaching set to the rail library to coach the classifier then take a look at the effectiveness by looking at a set (Joachims, T. 1998).

2.10 Comparison of Spam Filtering

The text categorization problem makes researchers to propose many methods to deal with email classification and text categorization in general. These methods can be grouped into two categories which are statics methods that are based on pre-defined address list and dynamic methods which are based on contents of the email (Yu, B., & Xu, Z. B. 2008). This means filtering the words and sentences, and then group spam and non-spam messages. The comparison of the email filtering methods are very important because they can help the one who wants to use the method to have a good choice. In recent years many researchers

have tried to write about comparison of methods and come out with many suggestions. The Naïve Bayes has been singled out as an effective method to construct automatically anti-spam filtering with good performance. In his paper, (Ion Androutsopoulos 2000) compared two approaches, Naïve Bayesian algorithm which was also used in (Sahami 1998), and memory-based of TiMBL (Daelemans et. Al. 1999). These two classifiers gave the results that showed both approaches achieved very high classification accuracy and precision.

Koutsias, J., et. al. (2000, July) two classifiers compared Naïve Bayesian with keyword-based anti-spam and their result shows that Naïve Bayesian performed much batter then the keyword based. The comparison of four classification on (Yu, B., & Xu, Z. B. 2008) which are Naïve Bayes (NB), Neural Network (NN), Support Vector Machine (SVM) and Relevance Vector Machine (RVM) shows that NN can achieve high accuracy compared to symbolic classifiers, only that it needs extensive time to select parameter and network training (Yu B and Z. B. 2008). The researchers (Hmeidi, I., et. la. 2015) by using Arabic dataset they tried to compare the classifiers which are Naïve Bayes, Support Vector Machine, Decision Tree, Decision Table, and K-Nearest Neighbor (KNN). The results showed that SVM leave behind all the other classifiers. According to these references which were mentioned earlier, we prefer to select among them those that many researchers have identified as the ones showing good performance when they are compared with other classifiers.

2.11 Dimensionality Reduction

When dealing with textual data there are some methods which have been developed to deal with this area. These have been divided in two groups, the supervised and unsupervised methods. Supervised methods use a set of pre-classified documents and consider the labeling of data. This means each text belong to one limited number of class and have a label which shows that text belong to which class, (Verbeek J. 2000), while unsupervised does not use the label. The feature is a group of attributes, in other words known as keywords that capture important data characteristics in a dataset. In feature selection, if the features for classifying is properly selected, then obviously the expected result

will be good, otherwise the result will not be as good as expected. This indicates that you must be careful on the selection method you choose.

2.11.1 Feature Extraction

Feature extraction is a process of making subset or new subset (dimensionality reduction). The main objective of Feature Extraction is to convert the free text view sentences into a set of words without losing their semantic meaning, (S. Vidhya 2007). The Feature extraction form a new set of the features by deducting some feature and make them small that will simplify the classification process. Feature extraction are used in machine learning when some methods are applied. In email filtering this helps to speed up the classification in text categorization in methods like SVM. Feature Extraction determine the words/terms that appear in a dataset/document.

The procedure explained by many researchers are PCA, ICA, Maximization of Mutual Information and a new variant of PCA that is called "Supervised-PCA" (S-PCA).

Term Frequency- Inverse Document Frequency (TF-IDF)

Term Frequency-IDF (TF-IDF) shows how many times a word appears in a document. The term frequency tf_{td} of term t (word) in a document d (dataset) is defined as the number of times that t occurs in d. relevance does not grow correspondingly with term frequency it may increase but not necessary proportional to term frequency, if the word appeared in a document 120 times this does not mean the relevance will be 120.

$$W_{td} = \log\left(1 + tf_{t,d}\right) * \log_{10}\left(\frac{N}{df_t}\right) \qquad (2.2)$$

The purpose of the method is to invention the illustration of the value of each document from a training dataset in which a vector between documents with terms will be established which then for the similarity among documents with the cluster will be determined by a prototype Vector also called cluster centroid.

The value of TF-IDF increases proportionately along with the number of words that appeared in the document, but is compensated by the frequency of words present in the corpus, that helps to regulate the fact that some words appeared many times and commonly used than others.

Term Frequency (TF)

Term Frequency $ft_{t,d}$ means that the term (t) in a document that is presented by 'd', TF calculate the number of times which the word appears in a document.

Document frequency (DF)

Document frequency in this approach shows that the rare terms are more informative compared to frequent terms. The terms like "increase, line, high" can be found in many documents. The document with these common terms are considered to be more relevant compared to the ones with rare terms.

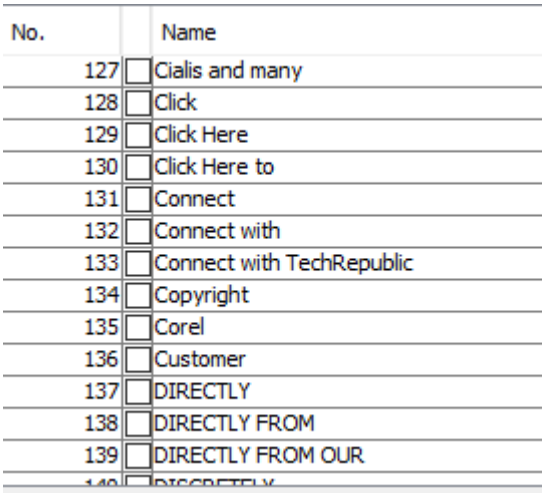
The Bag-of-Word (BoW) Model

The Bag-of-word model is among the mostly used feature extraction in spam filtering where the frequency of each word is used as a feature for training a classifier. This model does not consider order of words in a document. For example if one writes "Ally is running faster than Haji" or "Haji is running faster than Ally" these two sentences in BoW will be treated as same because it does not care about the grammar and even how your words structure in a sentence is formed but will be keeping multiplicity structure. The model describes documents by word frequency and totally ignore the relative position information of the words in the document. Bags can contain a repetition or redundant words. Some researchers said there are some specific strategies like Counting, Tokenization and normalization as bag of words.

WEKA Tokenization

The process of splitting up an arrangement of strings into occurrences such as phrases, keywords, words, symbols is called Tokenization process, there are many ways/methods for increasing the classifiers accuracy. In this research the author increases the accuracy by changing the tokenizers, in WEKA use three type of tokenizer which are:

- **WordTokenizer**: A simple tokenizer that is using the java.util.StringTokenizer class to tokenize the strings. The attributes for this include numbers '123', special characters '# &' and words 'hotel, administrator'. If there are two words that joined without space between them then, it is counted as one word 'AccorHotel'.
- **NGramTokenizer**: Splits a string into an n-gram with min and max grams. In the n-gram model, the technique is applied to characters or symbol but not like Word-Tokenizer which applied a word. The attribute for this tokenizer include alphabet "A", words "click" and also phrase, "click here", "click here to" as shown in picture 2.4 bellow:



No.	Name
127	Cialis and many
128	Click
129	Click Here
130	Click Here to
131	Connect
132	Connect with
133	Connect with TechRepublic
134	Copyright
135	Corel
136	Customer
137	DIRECTLY
138	DIRECTLY FROM
139	DIRECTLY FROM OUR

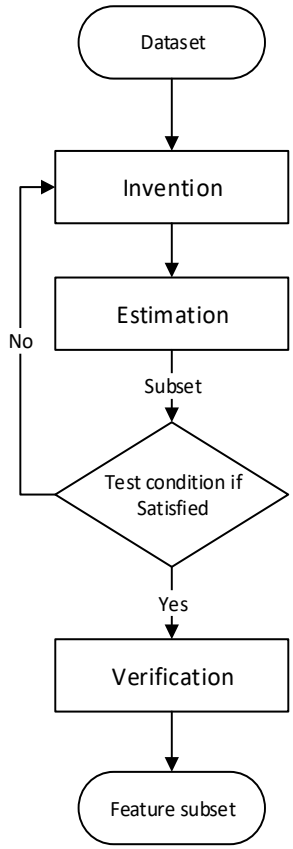
Picture 2. 4: NGram Tokenizer

- **AlphabeticTokenizer**: With Alphabetic string tokenizer, tokens are to be formed only from contiguous alphabetic sequences. The attributes for this tokenizer are alphabets and words only symbols special characters were not included.

2.11.2 Feature Selection

In machine learning and statistics the term feature selection (variable selection, attribute selection or variable subset selection) is a processes which helps to deduct some features in order to help the classifying process to work smoothly. The objective is to have good performance. (Kumar, N., & P, D. 2015). In Feature selection the main objective is to select the highly problem-related features and to eliminate excessive features. The excessive features involve noisy

and redundant features. (Hsu, H.H. and Hsieh, C.W., 2010). SVM are said to perform well and produce good results without employing any feature selection techniques (Aakanksha S. et la 2015). This means that some classifiers need help to give good results but for some that is not the case. Although there are basically many methods for accomplishing a feature selection process but they are classified in to two groups only, that is supervised and unsupervised.



Picture 2. 5: Process in Feature Selection

2.11.2.1 Supervised (Wrapper method)

Wrapper methods or predictive training uses selected subset of variables that allows to detect the possible interactions between variables and estimate error on select dataset. The use of a subset evaluator:-

- This will create all possible subset from the feature vector

- Then it will use a classification algorithm to induce classifier from the feature in each subset
- It will consider the subset of feature with which the classification algorism perform the best

To find a subset, the evaluator will use a search technique (random search, breadth first search, depth first search or a hybrid search)

2.11.2.2 Unsupervised (Filter method)

Filter method or experimental training using selected features. Look at input only. Select the subset that has the most information and use an attribute evaluator and ranker to rank all the features in your dataset. The number of features you want to select from your feature vector can always be defined. It also omits the features, one at a time that have a lower rank and show the predictive accuracy of your classification algorithm. Weights put by the ranker algorithm are different from those of classification algorithm.

2.12 Evaluation Measures

The confusion matrix is a technique for summarizing the performance of a classification algorithm, represented using a table that is many times used to explain on a set of tested data for which the true values are known. The number of predictions of accurate and inaccurate are sum up with count values and broken down by each class. The matrix is easy to read and understand. The only thing which is confusing are the terminology used. The confusion matrix demonstrates the ways in which classification model is confused when it makes predictions.

Table 2. 1: Confusion matrix

Filter		**Classified as**	
		Actual Positive	**Actual Negative**
Original	**Predictive Positive**	true positive (TP)	false positive (FP)
	Predict Negative	false negative (FN)	true negative (TN)

In our case of emails spam and non-spam, the value in confusion matrix are True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) this means that TP are the actual Non-spam that were correctly

classified as Non-spam, FP are the Spam that were imperfectly classified as Non-spam, FN Non-spam that were wrongly marked as Spam and the TN are Spam correctly classified as Spam.

Accuracy this is a percentage of value/data that is correctly classified from the total amount of data. The calculation formula is as follows.

$$\frac{TP + TN}{TP + FP + FN + TN}$$

Recall: This is a percentage of appropriate class that is correctly retrieved (TP) compared to all data in the same class (TP + FN). The formula is as follows.

$$\frac{TP}{TP + FN}$$

Precision is a percentage of appropriate data (class) that is correctly retrieved (TP) with regard to all classes retrieved from the same class (TP + FP). The calculation formula is as follows.

$$\frac{TP}{TP + FP}$$

Error Rate is an opposite of accuracy as we said before which means the percentage of data (class) that is wrongly classified.

$$\frac{FP + FN}{TP + FP + FN + TN}$$

F-Measure a global estimation of the performance that come by combining a single measure Precision (P) and Recall (R).

$$\frac{2PR}{R + P}$$

A classification system is declared effective if the calculation results show precision is high and does not qualify if the recall is low.

True Positive (TP) Rate – a rate of true positive that means instances which is correctly classified in a given class. This exposes the classifier's capability to detect instances of the positive class.

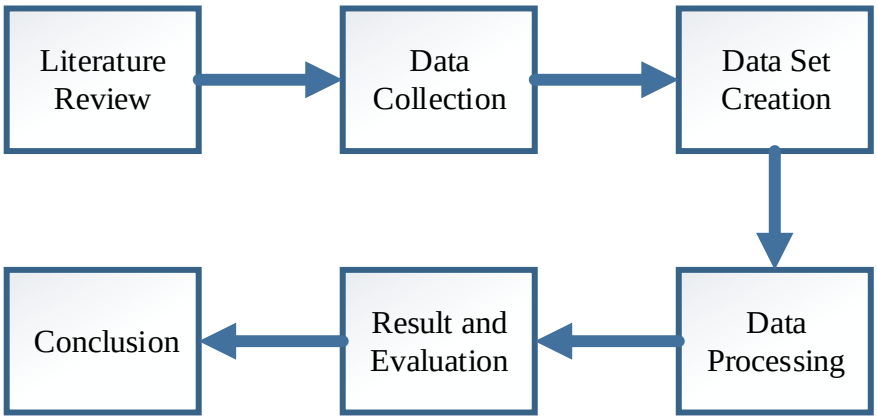
TPR = TP / (TP + FN)

False Positive (FP) Rate – a rate of false positive that means instances which is incorrect classified in a given class. This reflects the frequency with which the classifier makes a mistake by classifying normal state as pathological

$$FPR = FP / (FP + TN)$$

CHAPTER 3
RESEARCH METHODOLOGY

This chapter will explain the activities or steps which will be included when we conduct this research as shown in graph 3.1 and explain each step in detail. The steps start with literature review which will give us the idea on the topic. This will be followed by data collection which will explain how data were collected and where, and then a section on dataset creation will be discussed because in order to proceed we must have the dataset to train and to test the classifiers. Data processing including pre-processing, classifiers evaluation and result and conclusion will all be covered here.



Picture 3. 1: Research steps

3.1. Literature Review

The literature study is a very important activity that is used to get a comprehensive picture of what other researchers have done, how they did it and what is their argument. Also from that you can identify the research gap as the basis of research to be conducted. Previous studies serve as a support for researchers to conduct their new researches by identifying a problems in the field which in our case is email classification where we mainly base on machine learning techniques which is classifying them in two groups of things - like spam

and non-spam. So we will check on how other literature has said about the topic, methodology and other methods used. We will also check the classifiers used in these papers and their performance that will prove if what we want to do is relevant. We will also look into the dimensionality reduction and how it helps the identified classifiers to perform.

3.2. Data Collection

The nature of the research data can be grouped in to two forms; those are (1) Primary data, which is the data collected by the researchers fresh from the field and for the first time, meaning not yet processed. Briefly, primary data is the original data. Primary data involves direct experience and observation and as distortions by other observers were avoided, this makes primary data reliable. (2) Secondary data is the data which has been collected by either researchers, analyzed and have already passed through statistical process. The method used to collect the secondary data is by going through documents that have been written by the researchers who collected the data, analyzed them and wrote the reports that others refer to. This is a very important task because we need these data to show the performance of the classifier (trained and test).

The nature of this research and the data used are collected by researcher from the different e-mail addresses, and can therefore be known as the primary data. Four email addresses were used, three owned by researcher himself including raomar7972@gmail.com, indorashid@gmail.com, rashidthesis@gmail.com and one is from a close friend. In all these emails addresses we just want to have data which will be enough for our thesis. It took a long time for the data to be collected from the researchers email addresses because we had to wait may be a week or in some instances even a month to receive spam and non-spam email. This is what forced us to use a friend's email address because it was full of emails where almost 80% were spam emails. This email address simplified our task. The collection of emails were taken from both folders, inbox and spam. This will make training and testing of the classifier to be smooth and will be easy to accomplish the task.

3.3. Creation of the Dataset

After the completion of data collection exercise, the creation of the dataset process started. This is also a very important and interesting task. A new dataset created will be used in our thesis for training the classifier and also part of that dataset will be used to test classifiers in machine learning by using WEKA. The dataset has just three features which are sender address, the email body and the email subject and the fourth one is the label of it (Spam and Non-spam). This task also consume more time in the research, the dataset content are email which is text only used to train and check the performance of the classifiers. Also the text in that dataset are a mix of two languages which are English language and Swahili language known as "lugha ya Kiswahili" the language from East Africa.

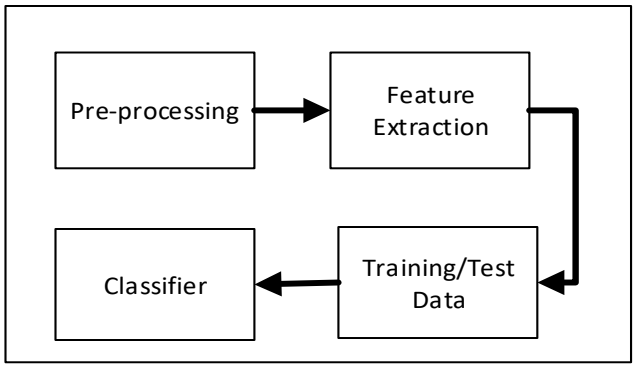
The process of dataset creation involve the data cleaning to make it suitable to run in the program which we plan to use. Generally, email messages contain many full stops, commas and quotation marks or single quotes which are not suitable to use when you run in WEKA. The input file format for WEKA system dataset is known as attribute-relation file format (ARFF). The example of dataset which we create the input file format which WEKA support for spam and non-spam data in that file structured like this have name of the relation (Email Spam n Non-spam) at the top, the block that define the attributes of features (sender, subject, body, Class {Spam, Non-spam}) Nominal attributes that is followed by the set of values can be enclosed in curly braces.

The researcher created his own dataset because no dataset of spam/non-spam which is written in Swahili language was found. This will be the first one to be created and it is believed that the move might encourage others to do so in order to make life a little bit easier for the upcoming researchers whose interest will be on Swahili language. The main challenge is in content of the data in the dataset as in the Swahili language we do not have many spam email messages so that we have been forced to use English to replace that.

3.4. Data Processing

An Email file is represented as a collection of feature vectors and defined as the word that belongs to feature vector (Joachims 1998). These vectors

together represent the number of email files and used to develop Term-Document Matrix (TDM). Usually, this matrix will be large and sparse in nature due to a large number of email files available for classification hence dimensionality reduction method is performed to tackle this problem that is done by feature selection and feature extraction processes. Some additional steps for dimension reduction of the matrix is also involved such as stop word (Least informative words such as pronouns, prepositions, and conjunction) removal (Joachims 1998) and lemmatization (grouping similar informative words such as Perform, Performed and Performing can be grouped as perform).



Picture 3. 2: Steps in Data Processing

3.4.1. Feature Extraction

The pre-processes are cleaning and tokenization. Both can be done using the StringToWordVector (STWV) filter in WEKA. To make the specified document able for classification using Machine Learning, we must do Feature extraction that convert a normal text to a set of features that ML Algorithm can understand and use to distinguish between spam and non-spam. In our case this will be done by STWV by assuming each word (String To Word) in the document is a feature and the number of occurrences in each instance is the feature value. In the research question one talk about feature extraction in order to make the work of classifier to be simple, so we will choose the best one which is widely used in many classification systems like TF-IDF to archive this goal.

3.4.2. Training and testing data

The first activity here is to use a training set in other word Supervised learning because the class label each tuple in the training set that means the classifier directed that the tuple belong to which class. In the creation of dataset we used to label the tuples each time a line of email was added so that the training dataset will work efficiently as planned. The performance measurement of the tuples which are already labeled in our case spam and non-spam. This will be compared with the WEKA application that will give the accuracy of the algorithm. This result will not affect the test set. The dataset which is planned to be used in this thesis have approximately 1000 tuples which will be labeled in to two different groups (spam and non-spam).

To forecast the performance of a classifier on a new data, we need to evaluate its error-rate on a dataset that participated in the structure of the classifier. This independent dataset is called the test set. The test set which can also be called unsupervised learning, the dataset will be smaller in number of tuples compared to the training set. Also when the tuples not classified that means not label in advance, the classifier will identify and group the tuples. The performance of the test set can be measured by checking whether it is correct or not. If it is correct, it is counted as success and otherwise that means error (error rate). Error rate as defined by Ian H. W 2005 "The error rate is just the proportion of errors made over a whole set of instances, and it measures the overall performance of the classifier". We assume that both the training data and the test data are representative samples of the underlying problem. The plan is to have a test set with not less than three hundred tuples for testing our classifiers.

3.4.3. Classification

The main activity in the classification is to filter the messages using classifier algorthm. In this research three classifier will be trained and tested with the dataset that was created before. The classifiers that many researchers agreed on their high perfomance, the Naïve Bayes is among the classifier that will be used in this research. We believe it will give good result. The other two are The Sequential Minimal Optimization (SMO) and The k-nearest neighbor K-NN.

False Negative can be reduced by applying very strong classifiers with the help of best feature extraction. Some of the work which have recently been published have proposed the idea of a partial Naive Bayes approach, influenced towards low false positive rates.

3.5. Result and Evaluation

The experiment results of the classifiers that involve the document labeled with a set of categories “training set” and the one which the performance measure calculated “test set” will be evaluated by the experiment.

3.5.1. Performance Measures

The performance measures which are well known in Text Categorization are measures of recall and precision. We too decided to use them. A recall for a category is defined as the percentage of correctly classified documents among all documents belonging to that category, and precision is the percentage of correctly classified documents among all documents that were assigned to the category by the classifier.

3.5.2. Comparison among Classifiers

The experiment results of classifiers will be analyzed, separated and compared among them that will give us the best classifier which can also be compared with the classifier which Gmail use to filter the electronic mails.

CHAPTER 4
PRELIMINARY PROCESSES

This chapter will explain how and where data have been collected, the process of creating dataset and the format, and pre-process of implementation of the dataset.

4.1. Data Collection

The data collection process started after reading many documents including journals. This process took place in four email addresses, three of them were the author's addresses, and the fourth one belonged to a close friend. The three email addresses owned by the researcher are raomar7972@gmail.com, indorashid@gmail.com, and rashidthesis@gmail.com, while the one belonging to a close friend is idhemblack87@gmail.com. It was from these email addresses that enough data for this thesis was collected. It took a long time for the data to be collected from the researchers email addresses because we had to wait for a week or in some instances even a month to receive spam and non-spam email. This is what forced us to use a friend's email address because it was full of emails where almost 80% were spam emails. This email address simplified our task. The collection of emails were taken from both folders, inbox and spam. The emails collected were about eight hundred, an amount that was enough for testing the classifiers. Swahili spam emails were very few compared to the English ones, because of the Swahili people still use the English language to create their spam.

4.2. Creation of the Dataset

The creation of a dataset is quite a tricky task, because dataset must be in the format that can be excecuted in machine learning tools. The input file format for WEKA system dataset is known as Attribute-Relation File Format (ARFF) was therefore used. The dataset has just four features which are sender address, the email body, the email subject and the fourth one is class which is labeled as Spam and Ham. This task also consumed a lot of time in this research, the dataset content are emails which are text only, this used to train and check the

performance of the classifiers. The researcher create two datasets, Swahili language dataset and English language dataset, after running them separately, researcher combine them to form one dataset that contents are mixed of English sand Swahili language.

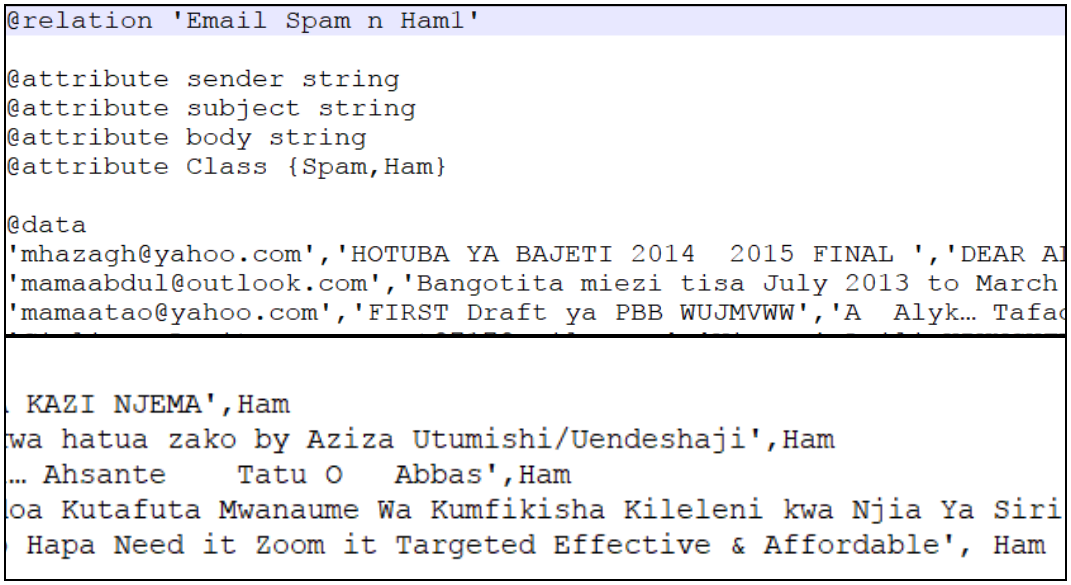
This process involve the data cleaning activity to make it suitable to run in the WEKA program. Generally, email messages contain many full stops (.), commas (,) and question mark (?) and single quotes (‘) which are not suitable to use in WEKA. Commas and single quotes (‘,’) are used in WEKA dataset as separator between attributes. So if the email message body contains one of them when executing the dataset, the error windows will appear with error line number.

Swahili dataset contains four hundred and fifty seven instances, among them four hundred and thirty eight instances are non-spam content and thirteen instances are spam content. English dataset contains four hundred and one instances, one hundred and eighty eight instances are spam content and two hundreds and thirteen instances are non-spam content. Combined English-Swahili dataset contains eight hundred and fifty eight instances, among them six hundred and fifty one instances are non-spam, and two hundred and seven instances are spam content. Statistical figures are shown in table 4.1 bellow:

Table 4. 1: Statistical Summary of Datasets

	Swahili Dataset		**English Dataset**		**Combined Dataset**	
	Spam	**Ham**	**Spam**	**Ham**	**Spam**	**Ham**
	19	438	188	213	207	651
Total	**457**		**401**		**858**	

The example of the dataset is shown in picture 4.1. The file structure has name of the relation (Email Spam n Ham1) at the top, the block that defines the features attributes (sender, subject, body, Class {Spam, Ham}) Nominal attributes that is followed by the set of values can be enclosed in curly braces, and also the instances.



Picture 4. 1: WEKA Dataset

4.3. Data Processing

An Email file is represented as a collection of feature vectors and defined as the word that belongs to feature vector (Joachims 1998). These vectors together represent the number of email files and used to develop Term-Document Matrix (TDM). Loading dataset which in this thesis are English dataset, Swahili dataset and combined dataset. Data pre-processing steps include the selection of filters, this research author used StringtoWordVector filter which find to be suitable for email messages, its transforms string attributes into a set of attributes representing word existence information from the text contained in the strings this depends on the tokenizer, in this research author used the default setting including tokenizer, WEKA in StingtoWordVector filter has three tokenizer the default is WordTokenizer, and other two are AlphabeticTokenizer and NGramTokenizer.

The pre-processing also including normalizing tokenized words, remove predefined set of words (stop-words). Stop words are the most common words in the Language, so filter help to omit them in order to speed up the search process by the help of dictionary. Stimming processing also the default one is used which is NullStemmer. After that next step are selecting the classify tab in order to select the classifiers, in this thesis three classifiers were used, which are SMO, Naïve

Bayes and k-NN. The next step author choose a classifier and run them one at a time. The author choose the 10-fold Cross-validation.

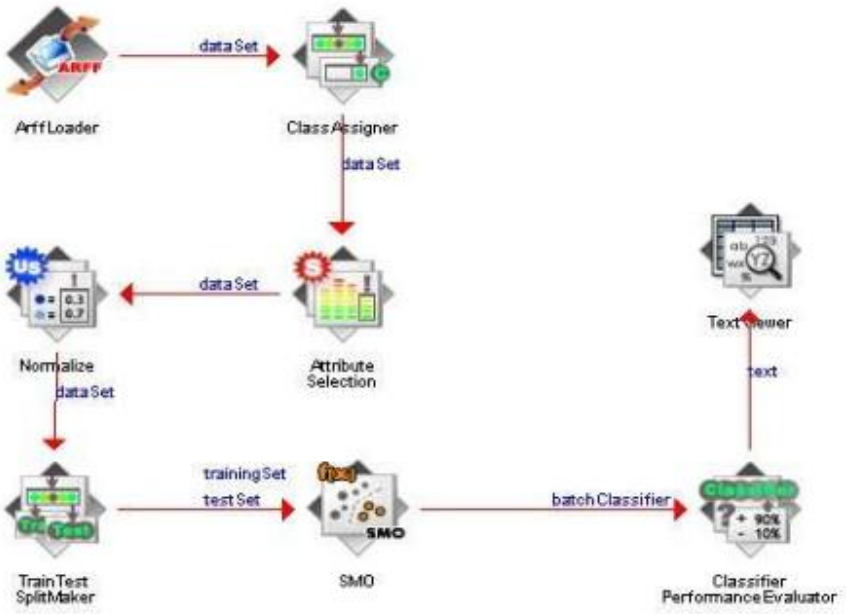
Cross-validation is a method of evaluating the predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. This means in this thesis the training set will be 70% and testing set will be 30% of instances of the dataset. All three dataset same setting was used.

CHAPTER 5
RESULTS AND EVALUATION

In this chapter, the author will display and discuss about the result obtained after doing the experiment of the created datasets in machine learning (WEKA).

This experiment gives the answer for the two questions that are written in chapter one, which are:

- What would be the performance of classifier if the dataset is a combination of two languages (Swahili Language and English Language)?
- How can the features be extracted in such a way that the classifiers' work could be simplified in order to increase accuracy?



Picture 5. 1: WEKA Processes in classification

5.1. The Results

In this experiment three dataset are created one from English language and one from Swahili language, then, experiment is conducted for these two, and

after getting the result, the researcher combine them to form one dataset and do experiment again. The classifiers used are Naïve Bayes, Sequential Minimal Optimization (SMO) and k-Nearest Neighbour (k-NN). The Researcher use StringToWordVector attribute filter to convert string to vector and 10 fold cross-validation is chosen as a test mode. Also, confusion matrix is used to predict and summarize the performance of a classifiers.

The confusion matrix is a technique for summarizing the performance of a classification algorithm. Table 5.1 is an example of it. The matrix is easy to read and understand. The confusion matrix demonstrates the ways in which classification model is confused when it makes predictions. Accuracy of classifier is calculated by (TP+TN)/total if these numbers are high that means the accuracy is good. Opposite is Misclassification Rate "Error Rate" which is calculated by (FP+FN)/total, if Error Rate is high that means bad classifier.

Table 5. 1 Confusion Matrix

	Actual Spam	**Actual Ham**
Predictive Spam	True Positive (TP)	False Positive (FP)
Predict Ham	False Negative (FN)	True Negative (TN)

5.1.1. English dataset

The English dataset contain four hundred and one instances, one hundred and eighty eight instances are spam content and two hundreds and thirteen instances are non-spam content.

5.1.1.1. Performance Measures for English Dataset

The English language dataset experimented by using three different algorisms that are SMO, Naïve Bayes and k-NN. Table 5.2 below “Confusion Matrix for English Dataset” shows correctly classified by number of emails where SMO has 391 correctly classified out of 401, Naïve Bayes 352 classified correctly out of 401, and k-NN 375 classified correctly out of 401. Percentage wise shown in Table 5.3 shows that SMO classifier leads other algorism, it achieved 97.51%,

followed by k-NN with average accuracy of 93.52% and the last but also good accuracy is Naïve Bayes that come with 87.78%.

Table 5. 2: Confusion Matrix for English Dataset

	CLASSIFIERS					
	SMO		**Naïve Bayes**		**k-NN**	
	Spam	**Ham**	**Spam**	**Ham**	**Spam**	**Ham**
Spam	182	6	144	44	171	17
Ham	4	209	5	208	9	204

The average from Precision, Recall, F-Measure and instances falsely classified as a given class Receiver Operator Characteristics (ROC) Area. SMO classifier still performs well on this by having average of Precision 0.975, Recall 0.975, F-Measure 0.975 and ROC area 0.975, followed by k-NN classifier which has Precision 0.936, Recall 0.935, F-Measure 0.935 and ROC area 0.939. The last one is Naïve Bayes classifier which come with Precision 0.892, Recall 0.878, F-Measure 0.876 and ROC area 0.98.

Table 5. 3: Average Classifiers performance

Classifier	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
SMO	**97.51%**	**0.975**	**0.975**	**0.975**	**0.975**
Naïve Bayes	87.78%	0.892	0.878	0.876	0.98
k-NN	93.52%	0.936	0.935	0.935	0.939

5.1.2. Swahili dataset

The Swahili dataset contains four hundred and fifty seven instances, among them four hundred and thirty eight instances are non-spam content and nineteen instances are spam content.

5.1.2.1. Performance Measures

The dataset experimented by using three different algorisms that are SMO, Naïve Bayes and k-NN. Table 5.4: “Confusion Matrix for Swahili Dataset” shows correctly classified by number of emails, SMO classifier has 451 correctly

classified instances and 6 incorrect classified instances out of 457 instances, Naïve Bayes classifier 453 classified correctly instances and 4 incorrect classified instances out of 457 instances, and k-NN classifier has 452 classified correctly instances and 7 incorrect classified instances out of 457 instances.

Table 5. 4: Confusion Matrix for Swahili Dataset

	CLASSIFIERS					
	SMO		**Naïve Bayes**		**k-NN**	
	Spam	**Ham**	**Spam**	**Ham**	**Spam**	**Ham**
Spam	13	6	15	4	12	7
Ham	0	438	0	438	0	438

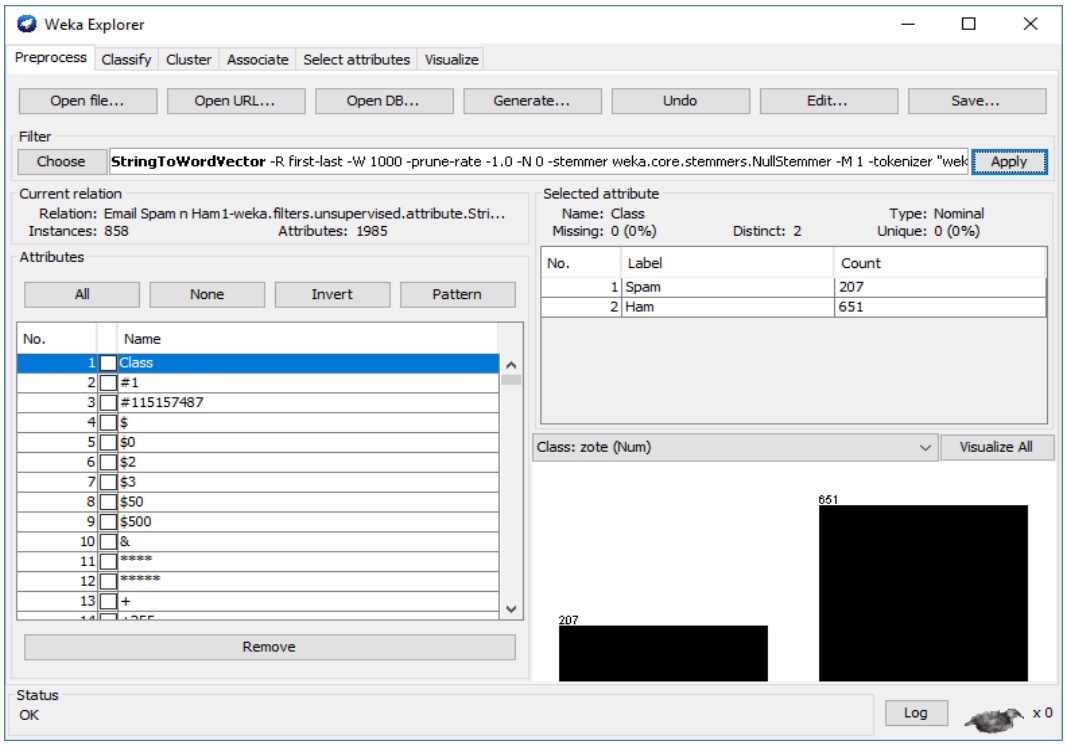
Percentage wise as shown in Table 5.5 indicates that the Naïve Bayes classifier leads other classifiers by 0.43%, it has correctly classified of 99.12%, which is followed by SMO classifier with average accuracy of 98.69%. SMO classifier lead k-NN by 0.22% and the last but also good accuracy is k-NN that come with 98.47% correctly classified. This can be concluded that the Naïve Bayes can work better with the Swahili language, although the gap from one classifier to another is not that big. Also table 5.5 continue to shows average of Precision, Recall, F-Measure and ROC Area. Naïve Bayes classifier still performs well on this by having average of Precision 0.991, Recall 0.991 Recall, F-Measure 0.991 and ROC area 1, followed by SMO classifier which has Precision 0.987, Recall 0.987, F-Measure 0.986 and ROC area 0.842. The last one is k-NN classifier which comes with Precision 0.985, Recall 0.985, F-Measure 0.983 and ROC area 0.873.

Table 5. 5: Average Classifiers accuracy by classes

Classifier	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
SMO	98.69%	0.987	0.987	0.986	0.842
Naïve Bayes	**99.12%**	**0.991**	**0.991**	**0.991**	**1**
k-NN	98.47%	0.985	0.985	0.983	0.873

5.1.3. Combined English and Swahili Dataset

The combined English-Swahili dataset contains eight hundred and fifty eight instances, among them six hundred and fifty one instances are non-spam, and two hundred and seven instances are spam content. The experiment also use StringToWordVector attribute filter used to modify datasets in a systematic fashion that means with string to vector, there are 858 instances and 1985 attributes in combined dataset. In picture 5.2 “WEKA Explorer” shows this in details.



Picture 5. 2: WEKA Explorer

5.1.3.1. Performance Measures

The dataset is experimented by using three different classifiers that are SMO, Naïve Bayes and k-NN. Table 5.6 “Confusion Matrix for Combined Dataset” for combined dataset shows correctly classified by number of instances, SMO classifier has 846 correctly classified instances and 12 incorrect classified instances out of 858 instances, Naïve Bayes classifier has 797 instances classified correctly and 61 incorrect classified instances out of 858 instances, and k-NN classifier has 834 instances classified correctly and 24 instances incorrect

classified out of 858 instances. The Gmail classifiers has 157 correctly classified out of 182 and 25 incorrect classified.

Table 5. 6: Confusion Matrix for Combined Dataset

	CLASSIFIERS							
	SMO		**Naïve Bayes**		**k-NN**		**Gmail**	
	Spam	**Ham**	**Spam**	**Ham**	**Spam**	**Ham**	**Spam**	**Ham**
Spam	198	9	166	41	190	17	56	23
Ham	3	648	20	631	7	644	2	101

The percentages shown in Table 5.7 “Average Accuracy of Classifiers”. The chart shows SMO classifier leads other classifiers, it achieved correctly classified 98.60%, followed by k-NN classifier which has average of 97.20% correctly classified, and last but also good accuracy is Naïve Bayes that comes with 92.89% correctly classified. Also Table 5.7 continue to show the average of Precision, Recall, F-Measure and ROC area. SMO classifier still perform well on this by having average of Precision 0.986, Recall 0.986, F-Measure 0.986 and ROC area 0.976, followed by k-NN classifier which has Precision 0.972, Recall 0.972, F-Measure 0.972 and ROC area 0.964, and Naïve Bayes classifier come with Precision 0.928, Recall 0.929, F-Measure 0.928 and ROC area 0.964. The Gmail correctly classified is 86.26%, average of Precision 0.6, Recall 0.96 and F-Measure 0.96

Table 5. 7: Average Classifiers Performance

Classifier	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
SMO	***98.60%***	***0.986***	***0.986***	***0.986***	***0.976***
Naïve Bayes	92.89%	0.928	0.929	0.928	0.956
k-NN	97.20%	0.972	0.972	0.972	0.964
Gmail	86.26%	0.71	0.96	0.74	

5.2. Increase Classifiers’ Accuracy

The accuracy of classifiers can be increased by selecting some features. In WEKA the accuracy can be increased by choosing the suitable filter for the

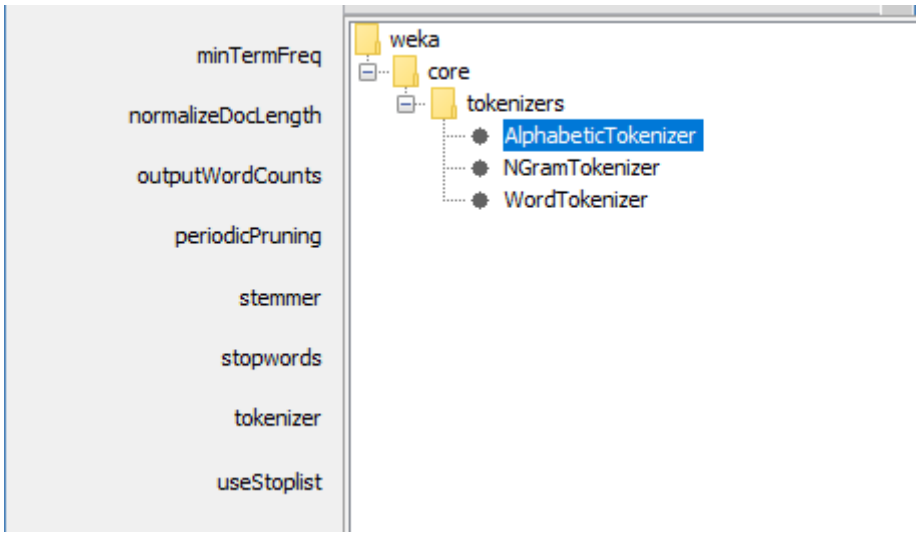
dataset which can help to bring about good result. Classifiers individually can be improved in different ways. The Naive Bayes classifier's performance is extremely sensitive to the selected attributes and the number of selected terms by the term-selection methods in the training stage (Almeida, T. A. et. el. 2011). Also accuracy can be increased by attribute subset selection, attribute creation and removing the redundant features (Kotsiantis, S. B., & Pintelas, P. E. 2004). However, it's not clear yet how features in email header can help to improve filtering result (Zhang, L., Zhu, J., & Yao, T. 2004).

SMO classifier which is an implementation of SVM use kernel function. It is well known that the choice of the kernel function is crucial to the efficiency of SVM. The four types of kernel functions are linear, polynomial, RBF and sigmoid frequently used with SVM. Yu, B., & Xu, Z. B. (2008) adopt sigmoidal kernel in the experiment, so the result shows that it does not matter whether it uses high volume of feature or low, the performance will remain the same. Joachims, T. (1998) use SVM with Polynomial and RBF kernels and compared with Naïve Bayes and k-NN, the performance of SVM was great. RBF also can be used in ANN (Kotsiantis, S. B., at. la. 2007). RBF is a three-layer feedback network. Individually, every unseen component implements a radial activation function and individually output component implements a weighted sum of hidden component outputs.

There are many methods applied for increasing the classifiers performance, in this research the author increases the accuracy by changing the tokenizers. Almeida, T. A., et. al. (2011, September) He tried to increase the performance by applying two types of tokenizers, The first one that targets to be viewed as a unit apart pattern, domain names and mail addresses by dots, this will help classifier to identify a domain even if subdomains are differ. The second type is a token that targets to identify symbols that are used in spam messages, so this will help in identifying the class of the message. Also in his further research he recommends to have standard tokenizers that can produce a bigger number of tokens and patterns to contribute to classifier abilities to separate no-spam messages from spams. In WEKA for example, StringtoWordVector filter uses three types of tokenizers which are:

- WordTokenizer: A simple tokenizer that is using the java.util.StringTokenizer sclass to tokenize the strings.
- NGramTokenizer: Splits a string into an n-gram with min and max grams. In this thesis the default setting for this minimum is 1 and maximum is 3.
- AlphabeticTokenizer: Alphabetic string tokenizer, tokens are to be formed only from contiguous alphabetic sequences.

In WEKA, STWV by default use Word Tokenizer, this tokenizer used in first experiment. The Author try to increase the classifiers' accuracy by using Alphabetic Tokenizer, so the results below was experimented by using STWV and in tokenizer the alphabetic tokenizer was selected. Picture 5.3 show list of tokenizers which are available in WEKA.



Picture 5. 3: WEKA list of Tokenizers

Table 5.14 confusion Matrix for SMO show results before that means Word-Tokenizer was used and after means the application of Alphabetic tokenizer. Before results shows TP 13 instances, TN 6 instances and FP 438 instances, no FN. After using the alphabetic tokenizer TP was 16 instances, TN 3 instances and FP 438 instances and no FN. This means the experiment identify three more instances.

Table 5. 8 : Confusion Matrix for SMO

	SMO CLASSIFIERS			
	Before		**After**	
	Spam	**Ham**	**Spam**	**Ham**
Spam	13	6	16	3
Ham	0	438	0	438

Table 5.9 shows the experimental results that shows the results before and after applying the alphabetic tokenizer. Before was 98.69% was correctly classified and with Precision 0.987, Recall 0.987, F-Measure 0.986 and ROC area 0.842, and after applying alphabetic tokenizer it shows 99.34% was correctly classified, the improvement was 0.65%, average of Precision 0.993, Recall 0.9933 and F-Measure 0.993 and ROC area is 0.921.

Table 5. 9: Average Details Performance for SMO

	Accuracy	**Precision**	**Recall**	**F-Measure**	**ROC Area**
Before	98.69%	0.987	0.987	0.986	0.842
After	**99.34%**	**0.993**	**0.993**	**0.993**	**0.921**

Combined dataset also its experimented by changing tokenizers this means all tokenizers were tested for this dataset to see if there is any impact if different tokenizers are used. The classifier used was SMO, and the three tokenizers used which are WordTokenizer, AlphabetiTokenizer and NGramTokenizer, minimum 1 and maximum 2 for N-gram tokenizer. Table 5.10 confusion matrix shows the results after experiment for combined dataset, NGramTokenizer got higher score compared to the WordTokenizer and AlphabetTekenizer, it has 848 correctly classified instances and 10 incorrect classified instances out of 858 instances. WordTokenizer and AlphabetTokenizer they got same results, both have 846 correctly classified instances and 12 incorrect classified instances out of 858 instances.

Table 5. 10: Confusion Matrix for SMO 'Tokenizers'

	SMO CLASSIFIERS					
	WordTokenizer		**AlphabetTokenize**		**NGramTokenizer**	
	Spam	**Ham**	**Spam**	**Ham**	**Spam**	**Ham**
Spam	198	9	198	9	199	8
Ham	3	648	3	648	2	649

SMO classifier performance after experiment for three tokenizers using combined dataset is shown in table 5.11. The results for NgramTokenizer was 98.48% accuracy, precision 0.988, recall 0.988, f-measure 0.988 and ROC area 0.979. WordTokenizer and AlphabetTokenizer come with same results, the performance was 98.60% accuracy and average of Precision 0.986, Recall 0.986, F-Measure 0.986 and ROC area 0.976.

Table 5. 11: Average Details Performance for SMO

Classifier	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
WordT	98.60%	0.986	0.986	0.986	0.976
AlphabetT	98.60%	0.986	0.986	0.986	0.976
NGramT	**98.83%**	**0.988**	**0.988**	**0.988**	**0.979**

5.3. Evaluation

The aim of this thesis was to have the dataset that combined two languages, English language and Swahili language. The dataset created was and experimented in WEKA by using three classifiers, which are Sequential Minimal Optimization (SMO) classifier which is implementation of SVM, Naïve Bayes classifier and k-NN classifier. But before combining that dataset, the researcher conducted experiment with separate English set and Swahili set.

The result in English dataset contains four hundred and one instances, among them one hundred and eighty eight instances are spam and two hundreds and thirteen instances are non-spam, and 2138 attributes. The English dataset result in accuracy for the classifiers shows that SMO classifier leads others. The

SMO has 391 correctly classified which is 97.51%, k-NN 375 classified correctly 93.52% and Naïve Bayes 352 classified correctly which is 87.78%.

The Swahili dataset contain four hundred and fifty instances among them four hundred and thirty eight instances are non-spam content and thirteen instances are spam content and contain 1686 attributes. The Swahili language emails messages for now they do not have many spam messages that means still Swahili emails can be trusted but they still use to circulate English spam messages in the area. The Swahili dataset result is different from English dataset where SMO leads other classifiers. Here, with the Swahili dataset Naïve Bayes leads other classifiers by having 453 classified correctly instances, in percentage this is 99.12%, followed closely by SMO classifier has 451 correctly classified instances equal to 98.69%, and last is k-NN classifier by having 452 classified correctly instances equal to 98.47%. The result above for Swahili dataset conforms to the result in (Ion Androutsopoulos 2000), (Sahami 1998), (Daelemans et. Al. 1999), (Koutsias, J., et. al. 2000, July) for Naïve Bayes to have good performance among other classifiers.

The combined English-Swahili dataset contains eight hundred and fifty eight instances, among them six hundred and fifty one instances are non-spam, and two hundred and seven instances are spam content. It has 1985 attributes. The experiment result for combined dataset shows SMO classifier leads other classifiers, it has 846 correctly classified instances this is equivalent to accuracy of 98.60%, followed by k-NN classifier which has 834 instances classified correctly and accuracy of 97.20%, and Naïve Bayes classifier which has 797 instances classified correctly with the accuracy of 92.89%. The average classifiers accuracy by classes, SMO classifier still perform well on this by having average of Precision 0.986, Recall 0.986 and F-Measure 0.986, followed by k-NN classifier which has Precision 0.972, Recall 0.972 and F-Measure 0.972, and Naïve Bayes classifier come with Precision 0.928, Recall 0.929 and F-Measure 0.928.

For both the English and the combined English – Swahili datasets, SMO implementation of SVM got best results because its support boundaries, also ability of execute very large dataset without requiring extra matrices storage, and it does not invoke any repetition of routine number for every sub-problem. This

results conforms with the one that was reported by (Al-Shargabi, B., Al-Romimah, W., & Olayah, F. 2011, April) and (Al-Kabi, M., Al-Shawakfa, E., & Alsmadi, I. 2013). SMO achieve higher than Naïve Bayes and J48 when researcher experiment them by using Arabic dataset, (Yu, B., & Xu, Z. B. 2008), and (Hmeidi, I., et. al. 2015) using Arabic dataset, the results showed that Support Vector Machine leave behind all the other classifiers. The combined dataset come with the results that the researcher predict to get in the proposal, which was to get high performance compared with the Gmail classifier. But things did not go well with the collection of emails. While it was expected that Swahili messages will have a large number of spam email, it was not so. Unfortunately the researcher found out that the Swahili language emails messages for now do not have many spam messages that means they still use English spam messages in the area.

Tokenizer changing has impact to classifier's performance as shown in results from experiment of combined dataset, and Swahili dataset. The classifier used was SMO classifier for all experiments and for all dataset (Swahili and Combined). In Swahili language dataset the author concentrated only in changing tokenizers, from word tokenizer to alphabet tokenizer. The results shown in table 5.9 had improved a little bit from accuracy of 98.69% to 99.34% this means increasing of 0.65%, and combined dataset three tokenizers were experimented, the results was N-gram tokenizer come with the good results after changing the maximum to 2 and minimum to 1, the default setting was maximum is 3 and minimum is 1, the accuracy for N-gram was 98.83%, the results was slightly deferent for two tokenizers (word and alphabet) both came with same results, in accuracy was 98.60%. This means that whether you choose word tokenizer or alphabet tokenizer there will be no change and their performance will be the same for this dataset.

The results of our experiment indicate that the combined dataset can give good results if N-gram tokenizer is used rather than Word tokenizer and Alphabet tokenizer. Krouska, A., Troussas, C., & Virvou, M. (2016, July) try to compare N-gram by changing from unigram, bigram and 1 to 3 grams classifier used were NB, SVM, k-NN and C4.5, the results shows 1-3 gram achieve good results for NB 92.59% which is higher for all classification experiment. The ability of n-

gram to detect word is higher some e-mails can have phrase like “n@ked l@dies” this can be extracted by n-grams as splitting words can be “n@k”, “@ked” (Goodman, J., Heckerman, D., & Rounthwaite, R. 2005) this can help to be identified as spam easily.

(This page is intentionally left blank)

CHAPTER 6
CONCLUSIONS AND RECOMMENDATIONS

This chapter will explains the conclusions of the research that has been conducted and suggestions to support further researches that may be possible to be carried out.

6.1. CONCLUSIONS

Swahili language is widely spoken in all East African countries for easy communication especially in the area of trade. The Swahili is complex somehow because it uses suffix on verbs, in directing the action against something. Also in negation it does not have specific words and position, sometimes can be in the beginning or at the end or in the middle of the verb. The Swahili standard is not like English language especially in vocabulary.

The Swahili emails are currently increasing in numbers and spread all over not only in East Africa but also in the world wherever the Swahili speakers travel to, work and reside. The precaution therefore must be taken and efforts needed to prevent before it is too late. If measures are not taken now, it will be very difficult in the next few years as Swahili people continue their study especially in new technology, this can make some of them to be bad guys that want to get money easily. This research will help the policy maker in East African countries to take this in to their considerations when they make ICT and Security Policies.

This research tried to answer two questions, first the performance of classifier if the dataset is a combination of two languages (Swahili Language and English Language). After the experiment the results show that SMO classifier has good performance in both the English dataset as well as the combined dataset, followed by k-NN classifier and Naïve Bayes classifier. Although Naïve Bayes classification result was not very good in English dataset and combined dataset, yet, it showed good performance for the dataset that was created by using Swahili language. This indicates that SMO classifier, k-NN classifier and Naïve Bayes all can be used in many languages, either by combining them or individually.

The second question was on the features to be extracted in such a way that the classifiers' work could be simplified in order to increase accuracy. The classification accuracy can be increased by selecting features, each classifier can use different ways to increase the performance. Some can be increased by select attribute, reduce redundant features, attribute subset selection, and attribute creation. SMO can be increased by choosing the kernel functions.

The algorithm that will fit to area that use mixed language like East Africa, because they also have two languages (English and Swahili) that are mostly used in the area not only at national level but also at international level as well. People used to compose or write their email by using those languages, sometimes they even mix them in one message. So the author recommend that when it comes to making decision as to which algorithm to use between SMO, Naïve Bayes and k-NN when they want to filter email messages, the answer is Sequential Minimal Optimization 'SMO'. It is the best choice for that because it was proved by the results in chapter 5, by achieving higher performance in combined dataset.

The experiment can have impact if tokenization settings are changed as shown in Chapter 5 when the author tried to change three tokenizers in String-to-Word-Vector, the results was deferent N-gram-tokenizer came with higher accuracy compared with word-tokenizer and alphabet-tokenizer.

6.2. RECOMMENDATIONS

Further research that might be possible to be conducted are to collect more Swahili language email messages especially Spam email messages and evaluate the result because very few Swahili spam email messages were collected in this research. The possibility of getting high performance by using your own filter is higher more than to use readymade, because it can be modified easily.

References

Aakanksha Sharaff et la (2015), Impact of Feature Selection Technique on Email Classification

Al-Kabi, M., Al-Shawakfa, E., & Alsmadi, I. (2013). The Effect of Stemming on Arabic Text Classification: An Empirical Study. Information Retrieval Methods for Multidisciplinary Applications, 207.

Almeida, T. A., Almeida, J., & Yamakami, A. (2011). Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers. Journal of Internet Services and Applications, 1(3), 183-200.

Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2011, September). Contributions to the study of SMS spam filtering: new collection and results. In Proceedings of the 11th ACM symposium on Document engineering (pp. 259-262). ACM.

Al-Shargabi, B., Al-Romimah, W., & Olayah, F. (2011, April). A comparative study for Arabic text classification algorithms based on stop words elimination. In Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications (p. 11). ACM.

Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., & Spyropoulos, C. D. (2000, July). An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 160-167). ACM.

Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000). Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. arXiv preprint cs/0009009.

Broomfield, G. (1930). The Development of the Swahili Language. Africa, 3(4), 516-522.

Contini-Morava, E. (2012). The message in the navel:(ir) realis and negation in Swahili. Language Sciences, 34(2), 200-215.

Goodman, J., Heckerman, D., & Rounthwaite, R. (2005). Stopping spam. Scientific American, 292(4), 42-49.

Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 10206-10222.

Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 10206-10222.

Hayati, P., & Potdar, V. (2008, November). Evaluation of spam detection and prevention frameworks for email and image spam: a state of art. In Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services (pp. 520-527). ACM.

Hmeidi, I., Al-Ayyoub, M., Abdulla, N. A., Almodawar, A. A., Abooraig, R., & Mahyoub, N. A. (2015). Automatic Arabic text categorization: A comprehensive comparative study. Journal of Information Science, 41(1), 114-124.

Hoanca, B. (2006). How good are our weapons in the spam wars? IEEE Technology and Society Magazine, 25(1), 22–30

Hsu, H.H. and Hsieh, C.W., (2010). Feature Selection via Correlation Coefficient Clustering. JSW, 5(12), pp.1371-1377.

Ian H. W and Eibe F. 2005 Data Mining “Practical Machine Learning Tools and Techniques” 2nd edition.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features (pp. 137-142). Springer Berlin Heidelberg.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98, 137-142.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. IEEE transactions on systems, man, and cybernetics, (4), 580-585.

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. Journal of advances in information technology, 1(1), 4-20.

Kotsiantis, S. B., & Pintelas, P. E. (2004, September). Increasing the classification accuracy of simple bayesian classifier. In International Conference on Artificial Intelligence: Methodology, Systems, and Applications (pp. 198-207). Springer, Berlin, Heidelberg.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

Krouska, A., Troussas, C., & Virvou, M. (2016, July). The effect of preprocessing techniques on Twitter sentiment analysis. In Information, Intelligence, Systems & Applications (IISA), 2016 7th International Conference on (pp. 1-5). IEEE.

Kumar, N., & P, D. (2015). Study on Feature Selection Methods for Text Mining.

Lam, H. Y., & Yeung, D. Y. (2007). A learning approach to spam detection based on social networks (Doctoral dissertation, Hong Kong University of Science and Technology).

Marjie-Okyere, S. M. (2013). Borrowings in Texts: A Case of Tanzanian Newspapers. New Media and Mass Communication, 16, 1-8.

Mojdeh, M. (2012). Personal Email Spam Filtering with Minimal User Interaction.

N pérez-díaz, & F fdez-riverola,. (2016). Boosting Accuracy of Classical Machine Learning Antispam Classifiers in Real Scenarios by Applying Rough Set Theory. Scientifc Programming, -(-), 1-11

Petzell, M., (2005). Expanding the Swahili vocabulary. Africa & Asia, 5, pp.85-107.

Phadke, S. G. (2015). Email Classification Using a Self-Learning Technique Based on User Preferences. Circulation, 701, 8888.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

Port, R. F. (1981). The applied suffix in Swahili. Studies in African Linguistics, 12(1), 71.

Saad, M. K. (2010). The impact of text preprocessing and term weighting on arabic text classification. Gaza: Computer Engineering, the Islamic University.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

Thota, H., Miriyala, R. N., Akula, S. P., Rao, K. M., Vellanki, C. S., Rao, A. A., & Gedela, S. (2009). Performance comparative in classification algorithms using real datasets. Journal of Computer Science and Systems Biology, 2(1), 97-100.

Tretyakov, K. (2004, May). Machine learning techniques in spam filtering. In Data Mining Problem-oriented Seminar, MTAT (Vol. 3, No. 177, pp. 60-79).

Tretyakov, K. (2004, May). Machine learning techniques in spam filtering. In Data Mining Problem-oriented Seminar, MTAT (Vol. 3, No. 177, pp. 60-79).

Verbeek, J., 2000, December. Supervised feature extraction for text categorization. In Tenth Belgian-Dutch Conference on Machine Learning (Benelearn'00).

Wang, L. (Ed.). (2005). Support vector machines: theory and applications (Vol. 177). Springer Science & Business Media.

Youn, S., & McLeod, D. (2007). A comparative study for email classification. Advances and innovations in systems, computing sciences and software engineering, 387-391.

Yu, B., & Xu, Z. B. (2008). A comparative study for content-based dynamic spam classification using four machine learning algorithms. Knowledge-Based Systems, 21(4), 355-362.

Zeng, Z. Q., Yu, H. B., Xu, H. R., Xie, Y. Q., & Gao, J. (2008, November). Fast training Support Vector Machines using parallel sequential minimal optimization. In Intelligent System and Knowledge Engineering, 2008. ISKE 2008. 3rd International Conference on (Vol. 1, pp. 997-1001). IEEE.

Zhang, L., Zhu, J., & Yao, T. (2004). An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing (TALIP), 3(4), 243-269.

Website

The Radicati Group www.radicati.com/

http://www.spamlaws.com/spam-stats.html

Anti-Phishing Working Group (APWG)

http://www.antiphishing.org/resources/apwg-reports/

The Spamhaus Project - The Definition of Spam

https://www.spamhaus.org/definition.html

https://www.pdx.edu/multicultural-topics-communication-sciences-disorders/swahili Portland state University

(This page is intentionally left blank)

THE AUTHOR'S BIOGRAPHIES

Rashid Abdullah Omar is among the four sons and one daughter of Mr. Abdullah's family. He was born in the Island of Zanzibar, which is part of the United Republic of Tanzania. The author acquired his Primary and High School education in Zanzibar. He studied his Bachelor degree in Computing and Information System at the Institute for Information Technology in Dar-es-salaam Tanzania and got second class with honors in 2006. From 2006 to 2010 he worked with the different companies in Dar-es-salaam. In May 2010 he went back to Zanzibar and worked in Government institutions. In 2015, the author joined the Institut teknologi Sepuluh Nopember Surabaya, Republic of Indonesia for his Masters in the same field of Information System and specialized in Security in the lab IKTI. The author successfully completed his postgraduate studies in March 2018.

The author's contacts are,
Email: indorashid@gmail.com
Facebook: Abuu Ibtisam Al-Zinjibari
LinkedIn: Rashid Omar
Line user ID: rasheedtanzania

(This page is intentionally left blank)